Supplementary Material: MAD Bayes for Tumor Heterogeneity – Feature Allocation with Exponential Family Sampling

A: Derivation of Equation (9)

Let $\phi(x) = x\log(\frac{x}{N}) + (N-x)\log(\frac{N-x}{N})$, we have $\nabla\phi(x) = \log(\frac{x}{N-x})$ and

$$\begin{aligned}\phi(n) - \phi(\mu) &= n\log(\frac{n}{N}) + (N-n)\log(\frac{N-n}{N}) - \mu\log(\frac{\mu}{N}) - (N-\mu)\log(\frac{N-\mu}{N}) \\ &= n\log(\frac{n}{\mu}) + (N-n)\log(\frac{N-n}{N-\mu}) + (n-\mu)\log(\frac{\mu}{N-\mu}).\end{aligned}$$

Therefore,

$$\begin{aligned}d_\phi(n,\mu) &= \phi(n) - \phi(\mu) - (n-\mu)\nabla\phi(\mu) \\ &= n\log(\frac{n}{\mu}) + (N-n)\log(\frac{N-n}{N-\mu}).\end{aligned}$$

The right-hand side of equation (9) is:

$$\begin{aligned}&\exp\{-d_\phi(n,\mu)\}f_\phi(n) \\ =\ &\exp\left\{n\log(\frac{\mu}{n}) + (N-n)\log(\frac{N-\mu}{N-n}) + n\log(\frac{n}{N}) + (N-n)\log(\frac{N-n}{N}) - h_1(n)\right\} \\ =\ &\exp\left\{n\log(\frac{\mu}{N}) + (N-n)\log(\frac{N-\mu}{N}) - h_1(n)\right\}.\end{aligned}$$

Given $\eta = \log(\frac{\mu}{N-\mu})$, the left-hand side of (9) is:

$$\begin{aligned}p(n \mid \mu) &= \exp\left\{n\log(\frac{\mu}{N-\mu}) - N\log(\frac{N}{N-\mu}) - h_1(n)\right\} \\ &= \exp\left\{n\log(\frac{\mu}{N-\mu}) - (N-n+n)\log(\frac{N}{N-\mu}) - h_1(n)\right\} \\ &= \exp\left\{n\log(\frac{\mu}{N}) + (N-n)\log(\frac{N-\mu}{N}) - h_1(n)\right\}.\end{aligned}$$

Thus equation (9) holds.

B: Derivation of (10)

$$
\begin{aligned}
L(\boldsymbol{Z},\boldsymbol{w}) &= p(\boldsymbol{Z})\, p(\boldsymbol{w} \mid \boldsymbol{Z})\, \widetilde{p}_\beta(\boldsymbol{n} \mid \boldsymbol{N},\boldsymbol{p}) \\
&= \frac{\gamma^C\, e^{-\gamma H_S}}{C!} \prod_{c=1}^{C} \frac{(S-m_c)!(m_c-1)!}{S!} \prod_{s=1}^{S} \left\{ \frac{\Gamma(\sum_{c=0}^{C} w_{sc})}{\sum_{c=0}^{C} \Gamma(w_{sc})} \prod_{c=0}^{C} w_{sc}^{a_c-1} \right\} \\
&\times \prod_{t=1}^{T}\prod_{s=1}^{S} \exp\left\{ -\beta \left[n_{st} \log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}}) \right] \right\} \\
&\times \exp\left\{ \beta n_{st} \left[\log(\frac{n_{st}}{N_{st}}) + (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}}) \right] - h_1(n_{st}) \right\}.
\end{aligned}
$$

Let $\gamma = \exp(-\beta\lambda^2)$ and consider $\beta \to \infty$, then

$$
\begin{aligned}
-\log L(\boldsymbol{Z},\boldsymbol{w}) &= \beta C\lambda^2 + \exp(-\beta\lambda^2) H_s + \beta \sum_{t=1}^{T}\sum_{s=1}^{S} \left[n_{st} \log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}}) \right. \\
&\quad \left. -n_{st} \log(\frac{n_{st}}{N_{st}}) - (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}}) \right] + O(1),
\end{aligned}
$$

where $u(\beta) = O(v(\beta))$ indicates there exist constants U_1 and U_2 such that $|u(\beta)| \le U_1 |v(\beta)|$ for all $\beta > U_2$.

It follows that

$$
\begin{aligned}
-\frac{1}{\beta}\log L(\boldsymbol{Z},\boldsymbol{w}) &= C\lambda^2 + O(\frac{\exp(-\beta\lambda^2)}{\beta}) + \sum_{t=1}^{T}\sum_{s=1}^{S} \left[n_{st} \log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}}) \right. \\
&\quad \left. -n_{st} \log(\frac{n_{st}}{N_{st}}) - (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}}) \right] + O(1/\beta) \\
&\sim C\lambda^2 + \sum_{t=1}^{T}\sum_{s=1}^{S} \left[n_{st} \log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}}) \right. \\
&\quad \left. -n_{st} \log(\frac{n_{st}}{N_{st}}) - (N_{st}-n_{st}) \log(\frac{N_{st}-n_{st}}{N_{st}}) \right] \\
&\sim C\lambda^2 + \sum_{t=1}^{T}\sum_{s=1}^{S} \left[-n_{st} \log(p_{st}) - (N_{st}-n_{st}) \log(1-p_{st}) \right]
\end{aligned}
$$

since $\frac{\exp(-\beta\lambda^2)}{\beta} \to 0$ and $1/\beta \to 0$ as $\beta \to \infty$.

Therefore, (10) holds.

C: Separable Convexity of the Objective Function Q

The objective function is

$$Q(\boldsymbol{p}) = \sum_{s=1}^{S}\sum_{t=1}^{T}\{-n_{st}\log(p_{st}) - (N_{st}-n_{st})\log(1-p_{st})\} + C\lambda^2.$$

Since $\log(x)$ and $\log(1-x)$ are both concave functions,
$\sum_{s=1}^{S}\sum_{t=1}^{T}\{-n_{st}\log(p_{st}) - (N_{st}-n_{st})\log(1-p_{st})\}$ is a convex function. Finally, since $C\lambda^2$ is a constant, the objective function $Q(\boldsymbol{p})$ is separable convex.

D: Proof of Theorem 3.1

Proof. By construction, in any iteration, the first and second steps do not increase the objective. We always choose the optimized values of $\boldsymbol{z}_s$ and $\boldsymbol{w}_t$ numerically.

Since no more than one feature is unique to any data point and no feature contains identical indices using left order form, the number of feature allocations is finite, which guarantees that the algorithm finishes in a finite number of iterations. □

E: Derivation of Small-variance Asymptotics to Modeling Subclone

$$\begin{aligned}
L(\widetilde{\boldsymbol{Z}}, \boldsymbol{w}) &= p(\widetilde{\boldsymbol{Z}})\, p(\boldsymbol{w} \mid \widetilde{\boldsymbol{Z}})\, p(\boldsymbol{\pi} \mid \widetilde{\boldsymbol{Z}})\, \widetilde{p}_{\beta}(\boldsymbol{n} \mid \boldsymbol{N}, \boldsymbol{p}) \\
&= \frac{\gamma^C\, e^{-\gamma H_S}}{C!}\prod_{c=1}^{C}\frac{(S-m_c)!(m_c-1)!}{S!}\pi_c^{m_{c1}}(1-\pi_c)^{m_c-m_{c1}}\prod_{s=1}^{S}\left\{\frac{\Gamma(\sum_{c=0}^{C}w_{sc})}{\sum_{c=0}^{C}\Gamma(w_{sc})}\prod_{c=0}^{C}w_{sc}^{a_c-1}\right\} \\
&\times \prod_{t=1}^{T}\prod_{s=1}^{S}\exp\left\{-\beta\left[n_{st}\log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}})\right]\right\} \\
&\times \exp\left\{\beta n_{st}\left[\log(\frac{n_{st}}{N_{st}}) + (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}})\right] - h_1(n_{st})\right\}.
\end{aligned}$$

Let $\gamma = \exp(-\beta\lambda^2)$ and consider $\beta \to \infty$, then

$$\begin{aligned}
-\log L(\widetilde{\boldsymbol{Z}}, \boldsymbol{w}) &= \beta C\lambda^2 + \exp(-\beta\lambda^2)H_s + \beta \sum_{t=1}^{T}\sum_{s=1}^{S}\Bigg[n_{st}\log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}}) \\
&\quad -n_{st}\log(\frac{n_{st}}{N_{st}}) - (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}})\Bigg] + O(1).
\end{aligned}$$

It follows that

$$\begin{aligned}
-\frac{1}{\beta}\log L(\widetilde{\boldsymbol{Z}}, \boldsymbol{w}) &= C\lambda^2 + O(\frac{\exp(-\beta\lambda^2)}{\beta}) + \sum_{t=1}^{T}\sum_{s=1}^{S}\Bigg[n_{st}\log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}}) \\
&\quad -n_{st}\log(\frac{n_{st}}{N_{st}}) - (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}})\Bigg] + O(1/\beta) \\
&\sim C\lambda^2 + \sum_{t=1}^{T}\sum_{s=1}^{S}\Bigg[n_{st}\log(\frac{n_{st}}{N_{st}p_{st}}) + (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}-N_{st}p_{st}}) \\
&\quad -n_{st}\log(\frac{n_{st}}{N_{st}}) - (N_{st}-n_{st})\log(\frac{N_{st}-n_{st}}{N_{st}})\Bigg] \\
&\sim C\lambda^2 + \sum_{t=1}^{T}\sum_{s=1}^{S}\Bigg[-n_{st}\log(p_{st}) - (N_{st}-n_{st})\log(1-p_{st})\Bigg]
\end{aligned}$$

since $\frac{\exp(-\beta\lambda^2)}{\beta} \to 0$ and $1/\beta \to 0$ as $\beta \to \infty$.

F: Details of MCMC

1. Updating $\boldsymbol{Z}$

   We update z_{sc} for $s = 1, \ldots, S$ and $c = 1, \ldots, \hat{C}$,

$$\begin{aligned}
\mathrm{p}(z_{sc} = 1|\text{ rest }) &\propto \frac{m_{-s,c}}{S}\prod_{t=1}^{T}\binom{N_{st}}{n_{st}}(p'_{st})^{n_{st}}(1-p'_{st})^{(N_{st}-n_{st})}, \\
\mathrm{p}(z_{sc} = 0|\text{ rest }) &\propto \frac{S-m_{-s,c}}{S}\prod_{t=1}^{T}\binom{N_{st}}{n_{st}}(p''_{st})^{n_{st}}(1-p''_{st})^{(N_{st}-n_{st})},
\end{aligned}$$

where

$$\begin{aligned} m_{-s,c} &= \sum_{s'=1|s'\neq s}^{S} z_{s'c}, \\ p'_{st} &= w_{t0}p_0 + \sum_{c'=1|c'\neq c}^{\hat{C}} w_{tc'}z_{sc'} + w_{tc}, \\ p''_{st} &= w_{t0}p_0 + \sum_{c'=1|c'\neq c}^{\hat{C}} w_{tc'}z_{sc'}. \end{aligned}$$

2. Updating $\boldsymbol{w}$

   To update $\boldsymbol{w}$, we assume $\tau_{tc} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a, 1)$ for $c = 1, \ldots, \hat{C}$ and $\tau_{t0} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a_0, 1)$. Define $w_{tc} = \tau_{tc} / \sum_{c'=0}^{\hat{C}} \tau_{tc'}$. This is equivalent to $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \text{Dir}(a_0, a, \ldots, a)$, $t = 1, \ldots, T$.

   (a) For $c = 0$

$$\text{p}(\tau_{t0}|\text{ rest }) \propto \tau_{t0}^{a_0-1}\exp(-\tau_{t0})\prod_{s=1}^{S} p_{st}^{n_{st}}(1-p_{st})^{(N_{st}-n_{st})}.$$

   We do the Metropolis-Hastings algorithm to update τ_{t0} on its logarithmic scale. Specifically, let $\tau'_{t0} = \exp(\log(\tau_{t0}) + u)$ where $u \sim \text{N}(0, \nu_\tau^2)$. That is, a random walk proposal in $\log(\tau_{t0})$. Then the acceptance probability is $\min(1, \rho_\tau)$ where

$$\rho_\tau = \frac{(\tau'_{t0})^{a_0}\exp(-\tau'_{t0})\prod_{s=1}^{S}(p'_{st})^{n_{st}}(1-p'_{st})^{(N_{st}-n_{st})}}{\tau_{t0}^{a_0}\exp(-\tau_{t0})\prod_{s=1}^{S} p_{st}^{n_{st}}(1-p_{st})^{(N_{st}-n_{st})}},$$

   where p_{st} and p'_{st} are evaluated with τ_{t0} and τ'_{t0}, respectively. Note that the Jacobian $= \tau_{t0}$.

   (b) For $t = 1, \ldots, T$ and $c = 1, \ldots, C$

$$\text{p}(\tau_{tc}|\text{ rest }) \propto \tau_{tc}^{a-1}\exp(-\tau_{tc})\prod_{s=1}^{S} p_{st}^{n_{st}}(1-p_{st})^{(N_{st}-n_{st})}.$$

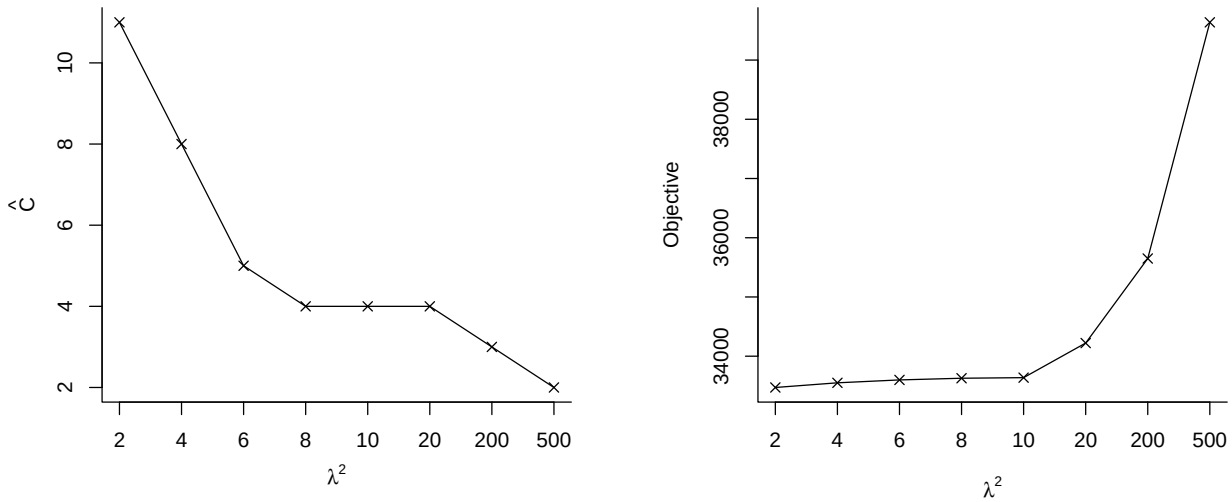
Let $\tau'_{tc} = \exp(\log(\tau_{tc}) + u)$ where $u \sim \mathrm{N}(0, \nu_\tau^2)$. Then the acceptance probability is $\min(1, \rho_\tau)$ where

$$\rho_\tau = \frac{(\tau'_{tc})^a \exp(-\tau'_{tc}) \prod_{s=1}^{S} (p'_{st})^{n_{st}} (1 - p'_{st})^{(N_{st} - n_{st})}}{\tau_{tc}^a \exp(-\tau_{tc}) \prod_{s=1}^{S} p_{st}^{n_{st}} (1 - p_{st})^{(N_{st} - n_{st})}},$$

where p_{st} and p'_{st} are with τ_{tc} and τ'_{tc}, respectively. The Jacobian $= \tau_{tc}$.

G: Supplementary Figures

Figure S1: Calibration Results.



(a) (b)

Figure S1: Calibration results. Panel (a) shows how the estimated number of features decreases with INCREASING λ^2. Panel (b) plots objective values as λ^2 increases.

Figure S2: Estimated Mean Cellular Prevalence of Each Cluster by PyClone

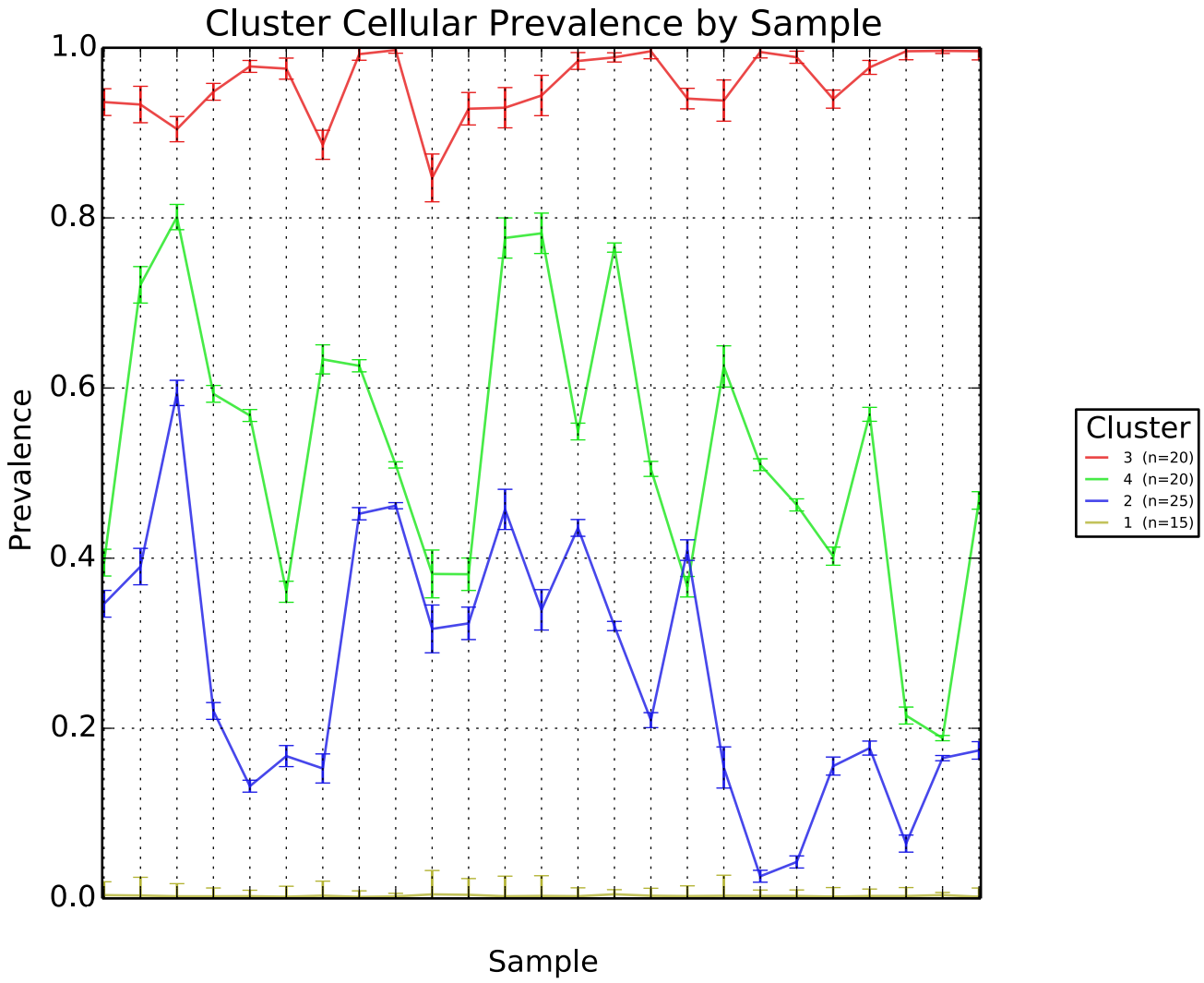
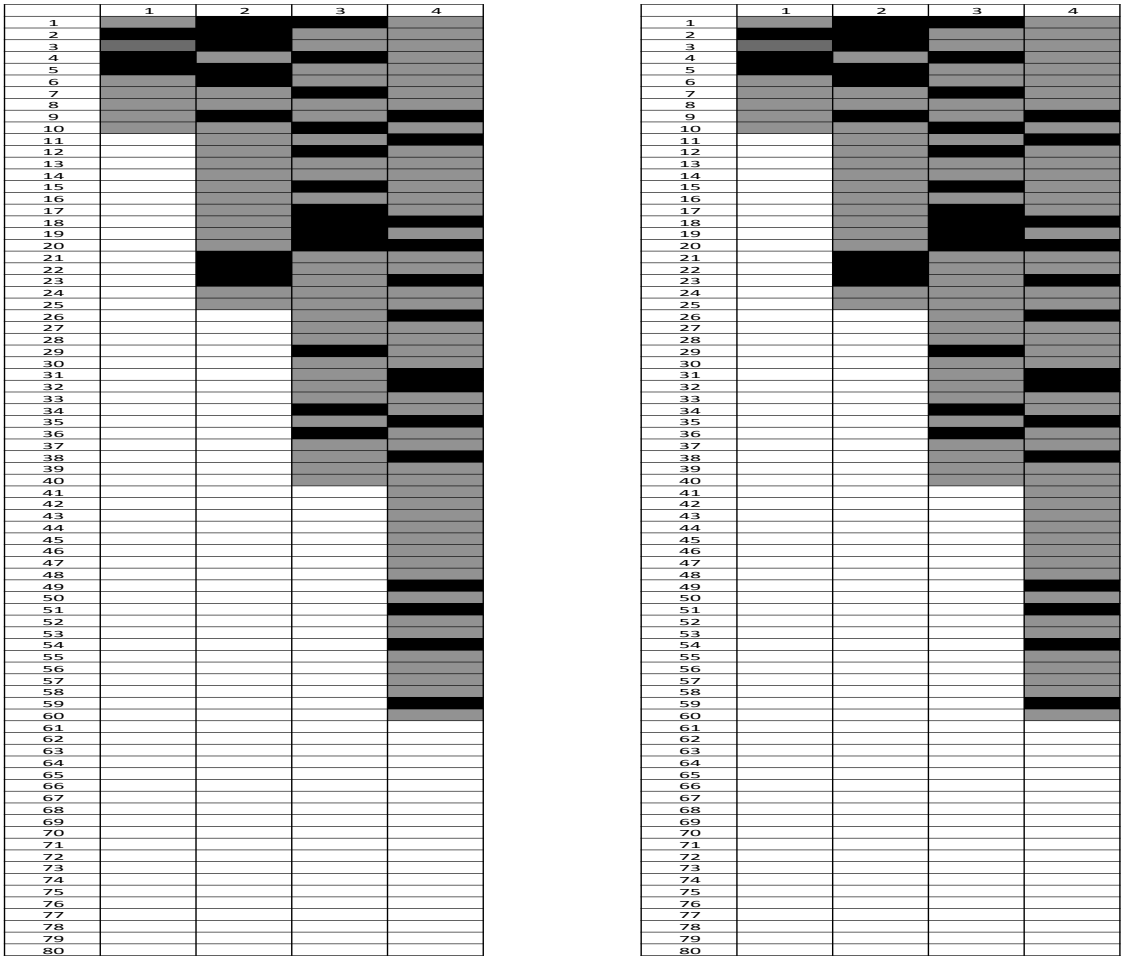


Figure S2: Estimated mean cellular prevalence of each cluster across all the 25 samples by PyClone.

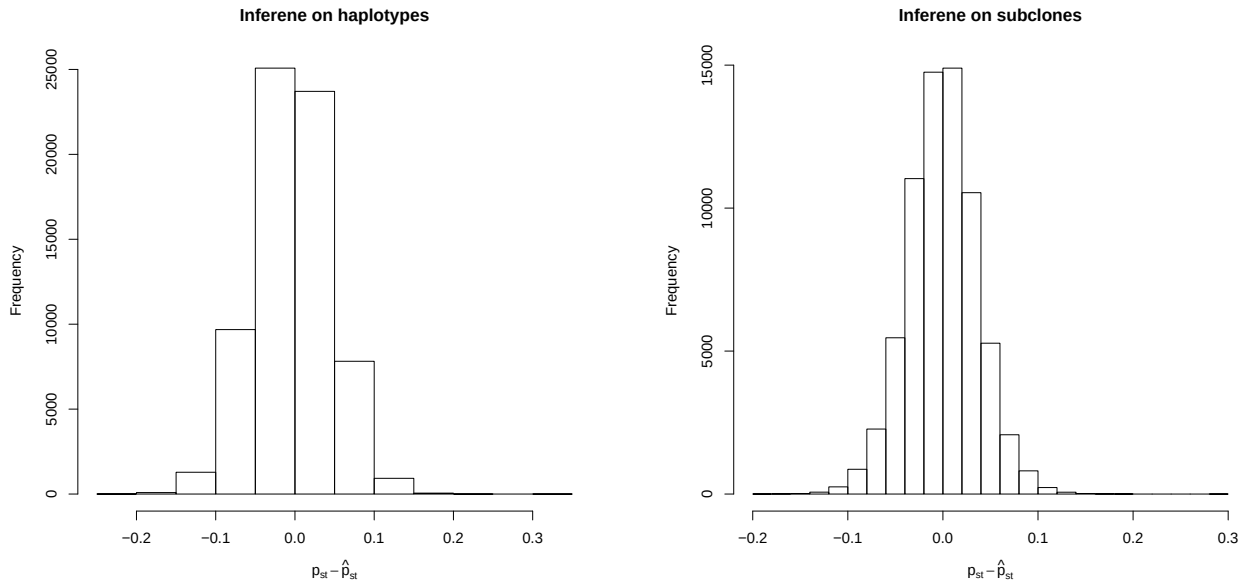
Figure S3: Simulation for Modeling Subclones



(a) Simulation Truth (b) Estimated (λ^2=50)

Figure S3: shown are feature allocation matrix $\widetilde{\boldsymbol{Z}}$, with grey shaded area indicating $\widetilde{z}_{sc} = 1$ and black shaded area indicating $\widetilde{z}_{sc} = 2$. Rows are SNVs and columns are inferred subclones. Panel (a) displays the simulation truth $\widetilde{\boldsymbol{Z}}^o$. Panel (b) displays the estimated feature allocation matrix $\hat{\widetilde{\boldsymbol{Z}}}$ when $\lambda^2 = 50$, which is exactly the same as the simulation truth.

Figure S4: Differences of $p_{st} - \hat{p}_{st}$



(a) (b)

Figure S4: The histogram of the differences of $p_{st} - \hat{p}_{st}$ in (a) haplotypes analysis, and (b) subclonal analysis.

Figure S5: Summary of PDAC Data

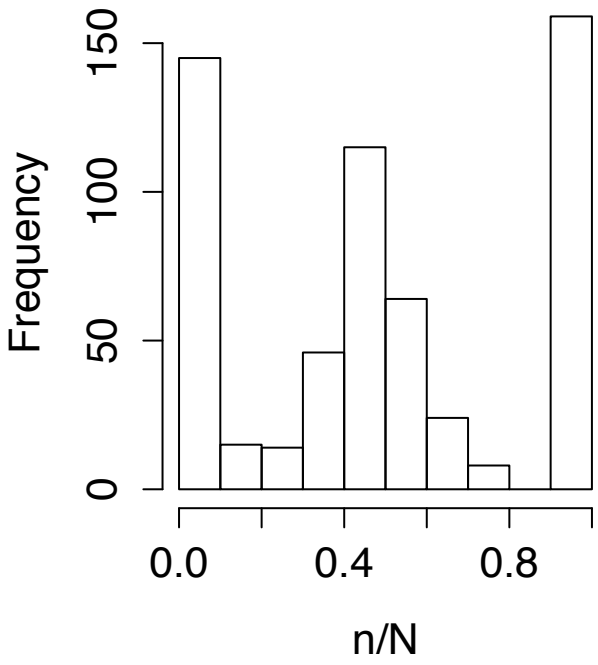


Figure S5: Summary of PDAC data.

Figure S6: Heatmaps of the Estimated Uncertainties for PDAC Data

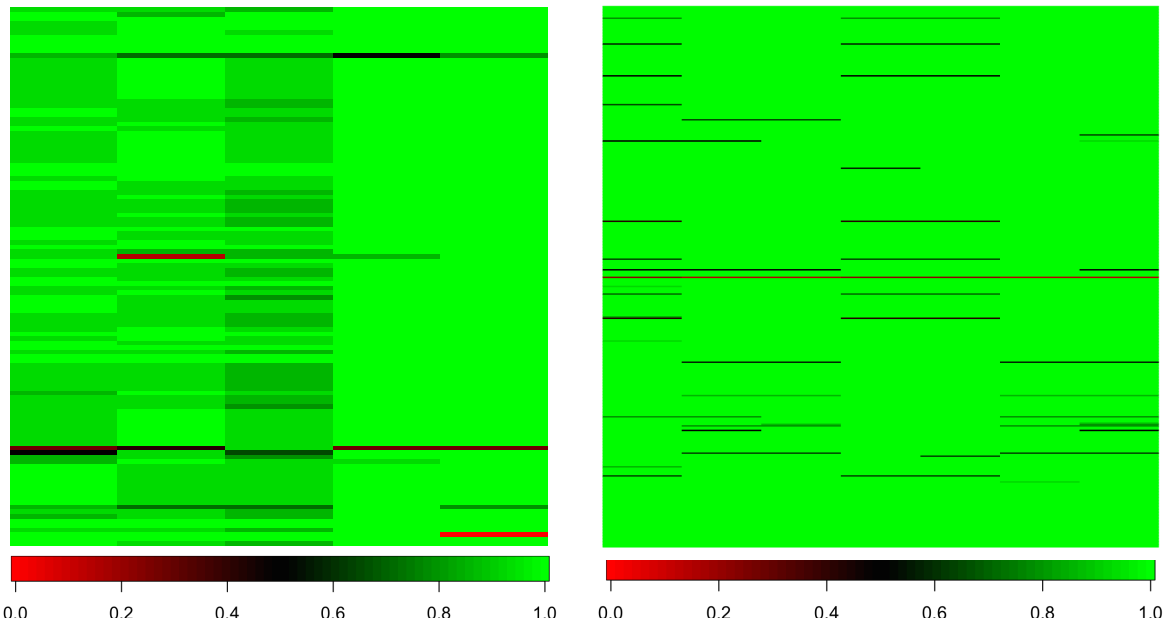


Figure S6: The heatmaps of the estimated uncertainties of (a) five estimated haplotypes for 118 SNVs using the PDAC data; (b) seven estimated haplotypes for 6,599 SNVs using the PDAC data.

H: Bioinformatics Data Processing for the Lung Cancer Data

Four surgically dissected tumor samples taken from the same patient with lung cancer were whole-exome sequenced. We extracted genomic DNA from each tissue and constructed an exome library from these DNA using Agilent SureSelect capture probes. The exome library was then sequenced in paired-end fashion on an Illumina HiSeq 2000 platform. About 60 million reads - each 100 bases long - were obtained. Since the SureSelect exome was about 50 Mega bases, raw (pre-mapping) coverage was about 120 fold. We then mapped the reads to the human genome (version HG19) (Church *et al.*, 2011) using BWA (Li and Durbin, 2009)

and called variants using GATK (McKenna *et al.*, 2010). Post-mapping, the mean coverage of the samples was between 60 and 70 fold.

A total of nearly 115,000 SNVs and small indels were called within the exome coordinates. We restricted our attention to SNVs that (i) make a difference to the protein translated from the gene, and (ii) that exhibit significant coverage in all samples with n_{st}/N_{st} not being too close to 0 or 1.

References

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R., *et al.* (2011). Modernizing reference genome assemblies. *PLoS biology* **9**, 7, e1001091.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* **25**, 14, 1754–1760.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research* **20**, 9, 1297–1303.